DOI:
10.1016/j.fsigen.2015.09.004

Document Version
Peer reviewed version

Title: Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans

Author: Ewelina Pośpiech Joanna Karłowska-Pik Magdalena Marcińska Sarah Abidi Jeppe Dyrberg Andersen Margreet van den Berge Ángel Carracedo Mayra Eduardoff Ana Freire-Aradas Niels Morling Titia Sijen Małgorzata Skowron Jens Söchtig Denise Syndercombe-Court Natalie Weiler The EUROFORGEN-NoE Consortium Peter M. Schneider David Ballard Claus Børsting Walther Parson Chris Phillips Wojciech Branicki

PII: S1872-4973(15)30066-1
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2015.09.004
Reference: FSIGEN 1410

To appear in: *Forensic Science International: Genetics*

Received date: 28-5-2015
Revised date: 10-8-2015
Accepted date: 9-9-2015

Please cite this article as: Ewelina Pośpiech, Joanna Karlowska-Pik, Magdalena Marcińska, Sarah Abidi, Jeppe Dyrberg Andersen, Margreet van den Berge, Ángel Carracedo, Mayra Eduardoff, Ana Freire-Aradas, Niels Morling, Titia Sijen, Malgorzata Skowron, Jens Söchtig, Denise Syndercombe-Court, Natalie Weiler, The EUROFORGEN-NoE Consortium, Peter M.Schneider, David Ballard, Claus Borsting, Walther Parson, Chris Phillips, Wojciech Branicki, Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans, Forensic Science International: Genetics http://dx.doi.org/10.1016/j.fsigen.2015.09.004

Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans.

Ewelina Pośpiech[1*], Joanna Karłowska-Pik[2], Magdalena Marcińska[3], Sarah Abidi[4], Jeppe Dyrberg Andersen[5], Margreet van den Berge[6], Ángel Carracedo[7,8], Mayra Eduardoff[9], Ana Freire-Aradas[7], Niels Morling[5], Titia Sijen[6], Małgorzata Skowron[10], Jens Söchtig[7], Denise Syndercombe-Court[4], Natalie Weiler[6], The EUROFORGEN-NoE Consortium; Peter M. Schneider[11], David Ballard[4], Claus Børsting[5], Walther Parson[9,12], Chris Phillips[7], Wojciech Branicki[1,3]

[1] Department of Genetics and Evolution, Jagiellonian University, Krakow, Poland
[2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
[3] Institute of Forensic Research, Section of Forensic Genetics, Krakow, Poland
[4] Faculty of Life Sciences, King's College, London, UK
[5] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[6] Department of Human Biological Traces, Netherlands Forensic Institute, The Hague, The Netherlands
[7] Forensic Genetics Unit, Institute of Forensic Sciences, Faculty of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain
[8] Genomic Medicine Group, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Institute of Health Carlos III, Spain
[9] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[10] Department of Dermatology, Medical College of Jagiellonian University, Krakow, Poland
[11] Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne, Germany
[12] Forensic Science Program, The Pennsylvania State University, University Park, Pennsylvania, USA

*corresponding author
E-mail – ewelina.pospiech@uj.edu.pl

Highlights:

• A replication study was made of SNPs most closely associated with hair morphology variation in Europeans (assigning straight, wavy and curly hair phenotypes).

• Analysis of 670 samples from seven European populations revealed the strongest association for rs11803731 in *TCHH*, rs7349332 in *WNT10A* and rs1268789 in *FRAS1*.

• Applying three different mathematical models to assess the predictive capacity of the SNPs indicated neural networks gives the best performing model to predict straight hair with high sensitivity and better specificity than logistic regression and CRT methods.

• The combined TTGGGG SNP genotype (rs11803731-rs7349332-rs1268789) was identified as the best predictor, giving greater than 80% probability of straight hair.

• The reported study is the first assessment of the forensic suitability of hair morphology as an externally visible characteristic. The results provide the basis for extended analyses of SNPs associated with this trait.

Abstract

DNA-based prediction of hair morphology, defined as straight, curly or wavy hair, could contribute to an improved description of an unknown offender and allow more accurate forensic reconstructions of physical appearance in the field of forensic DNA phenotyping. Differences in scalp hair morphology are significant at the worldwide scale and within Europe. The only genome-wide association study made to date revealed the Trichohyalin gene (*TCHH*) to be significantly associated with hair morphology in Europeans and reported weaker associations for *WNT10A* and *FRAS1* genes. We conducted a study that centered on six SNPs located in these three genes with a sample of 528 individuals from Poland. The predictive capacity of the candidate DNA variants was evaluated using logistic regression; classification and regression trees; and neural networks, by applying a 10-fold cross validation procedure. Additionally, an independent test set of 142 males from six European populations was used to verify performance of the developed prediction models. Our study confirmed association of rs11803731 (*TCHH)*, rs7349332 (*WNT10A)* and rs1268789 (*FRAS1*) SNPs with hair morphology. The combined genotype risk score for straight hair had an odds ratio of 2.7 and these predictors explained ~8.2% of the total variance. The selected three SNPs were found to predict straight hair with a high sensitivity but low specificity when a 10-fold cross validation procedure was applied and the best results were obtained using the neural networks approach (AUC=0.688, sensitivity=91.2%, specificity=23.0%). Application of the neural networks model with 65% probability threshold on an additional test set gave high sensitivity (81.4%) and improved specificity (50.0%) with a total of 78.7% correct calls, but a high non-classification rate (66.9%). The combined TTGGGG SNP genotype for rs11803731, rs7349332, rs1268789 (European frequency=4.5%) of all six straight hair-associated alleles was identified as the best predictor, giving >80% probability of straight hair. Finally, association testing of 44 SNPs previously identified to be associated with male pattern baldness revealed a suggestive association with hair morphology for rs4679955 on 3q25.1. The study results reported provide the starting point for the development of a predictive test for hair morphology in Europeans. More studies are now needed to discover additional determinants of hair morphology to improve the predictive accuracy of this trait in forensic analysis.

Keywords: Hair morphology; Forensic DNA phenotyping (FDP); *TCHH*; *WNT10A*; *FRAS1*; Neural networks

1. Introduction

Genetic prediction of traits that help define a person's appearance, known as Forensic DNA Phenotyping (FDP) significantly extends the range of DNA-based tools that can be used for intelligence. Together with gender determination, the inference of biogeographic ancestry (BGA) and age estimation, DNA-based prediction of externally visible characteristics (EVCs) can help to initiate a detailed description of an unknown individual in criminal cases without any suspects or entries in DNA profile databases [1-5]. Prediction of EVCs is also useful in anthropological studies or to add focus in the identification of missing persons. So far, eye colour has been the most thoroughly researched visible characteristic, with a marked success achieved for blue and brown eye colour prediction [e.g. 6-9] closely followed by genetic prediction of hair colour, with high predictive accuracy for red hair [10-12]. The characteristics of scalp hair collectively provide a very distinctive feature of human appearance, therefore these traits represent a very promising field of FDP. Not only prediction of hair colour, but also hair loss, hair greying and hair morphology are informative for the description of appearance. Recently, the initial development of predictive DNA tests for male pattern baldness was completed, showing that prediction is viable when supported with age estimation and inference of biogeographic ancestry [13].

Hair morphology is a highly conspicuous trait in humans that varies between populations of different biogeographic ancestry and is particularly variable among Europeans. African and Melanesian populations are characterized by the presence of tightly curled hair, Asians predominantly have straight hair which is also thicker than hair in Africans and Europeans, whereas in European populations ~45% of individuals have straight hair, ~40% have wavy hair and the remaining ~15% curly hair [14-15]. The morphology of hair has been shown to be highly heritable: reaching 85-95% heritability [16]. However, very little is known about genetic determinants beyond variation in hair textures. It has been suggested that hair straightness/curliness is programmed within the hair follicle and is determined biologically by the distribution and type of hair keratins as well as different cell types within the hair [17-19]. The structure of the inner root sheath (IRS) of the hair follicle and the types of keratins plus other proteins present are particularly important in molding and strengthening hair structure [20-22]. Recent studies conducted on East Asian populations have revealed two genes and coding single nucleotide polymorphisms (SNPs) within each, *EDAR* (rs3827760) and *FGFR2* (rs4752566) to

be the variants that are most strongly associated with thick and straight hair in East Asians [23-25]. The straight hair-associated G allele in rs3827760 is however almost completely absent in European and African populations, suggesting convergent evolution of straight hair by different pathways in Europeans and East Asians. So far, only one genome-wide association study (GWAS) has been conducted on European populations by Medland et al. in 2009, revealing the Trichohyalin gene (*TCHH*) on chromosome 1 (chr1) to be a major hair morphology gene [26]. Trichohyalin is highly expressed in the IRS layer of the hair follicle and provides mechanical strength by crosslinking both to itself and to other proteins during formation of the cornified cell envelope (CE) - the outermost layer of the hair follicle [20]. Four SNPs located within or near the *TCHH* gene were found to be statistically significant for hair morphology determination. These are rs17646946, rs11803731, rs4845418 and rs12130862, that in combination explain ~6% of the total variance of hair morphology in Europeans [26]. No other genes reached genome-wide significance in Medland's study. However, suggestive associations with hair morphology were also observed for *FRAS1* on chr4 (rs1268789) and *WNT10A* on chr2 (rs7349332) [26]. Interestingly, the associations of *TCHH* and *WNT10A* genes with hair morphology were further confirmed by Eriksson et al. in a study completed after that of Medland [27].

In the present study, we replicated data for six SNPs (rs17646946, rs11803731, rs4845418, rs12130862, rs1268789, rs7349332) in three genes (*TCHH*, *FRAS1* and W*NT10A*) identified by Medland [26], in a Polish sample consisting of 333 individuals with straight hair and 195 with curly/wavy hair. We assessed the predictive capacity of the six SNPs using three different mathematical methods of: logistic regression; neural networks; plus classification and regression trees (CRTs). Prediction models were evaluated using a 10-fold cross validation procedure. An additional set of 142 males from six other European populations was used to further assess the performance of the developed primary prediction models. Finally, analyses of association with hair morphology were also conducted for 44 SNPs previously identified to be associated with male pattern baldness (MPB), to investigate a possible link between the genetic determination of these two scalp hair traits in males: hair morphology and hair distribution.

2. Materials and Methods

2.1. Sample collection and phenotyping

Study samples were collected from 670 unrelated individuals from seven European populations: 528 samples from Poland (78.8%), 46 from the Netherlands (6.9%), 31 from the United Kingdom (4.6%), 26 from Denmark (3.9%), 19 from Italy (2.8%), 11 from Germany (1.6%) and 9 from Spain (1.4%). Hair morphology was assessed with a simple three-phenotype scale that categorized hair as straight, wavy or curly. Hair types corresponded to the previously proposed *Loussouarn* hair structure classification system [14], comprising: categories I and II for straight hair; category III for wavy hair; categories IV and V for curly hair. Phenotyping was conducted by direct inspection of study participants by a dermatology specialist combined with interview (Polish samples) or evaluation of high quality photographs of head hair (all other samples). All participating donors gave informed consent and the study was approved by each regional bioethics committee: the Commission on Bioethics of the Regional Board of Medical Doctors in Krakow (48 KBL/OIL/2008); the Danish Ethical Committee (H-3-2012-023); the KCL BDM Research Ethics Subcommittee (BDM/13/14-111); the USC ethics committee of clinical investigation in Galicia, Spain (CEIC: 2009/246) and followed approved internal procedures for the Netherlands Forensic Institute in The Hague.

2.2. SNP selection and genotyping

Six autosomal SNPs were selected for genotyping (rs17646946, rs11803731, rs4845418, rs12130862, rs1268789 and rs7349332) located in the genes: *TCHH*, *FRAS1* and W*NT10A*; all associated with hair morphology from the only GWA study made of this trait to date [26]. Heilmann et al. recently reported an association of rs7349332 in *WNT10A* with male pattern baldness [28]. This finding indicated a possible link between genetic susceptibility to human hair loss and hair morphology. In order to address this issue we investigated a set of 44 SNPs, including 23 autosomal and 21 X-linked SNPs, previously reported to be associated with male pattern baldness [28-36]. Detailed information on all the analyzed polymorphisms is given in Supplementary Table S1. DNA extraction and genotyping was performed as described previously by Marcińska et al. [13]. Briefly, for most samples, SNPs were genotyped in four single base extension (SBE) assays using the SNaPshot system (Applied Biosystems, Foster City, CA, USA). All details, including reactions conditions, PCR and SBE primers sequences are as described in Marcińska et al. [13]. The UK samples were genotyped with next-generation sequencing using the Illumina MiSeq system. For this purpose, DNA was amplified under identical conditions to

the four PCR reactions described above using the same PCR primers. PCR products were then combined, quantified using the Qubit dsDNA high-sensitivity assay kit (Life Technologies) and 50ng of this double stranded product-set was then taken through to the library preparation stage. DNA libraries were constructed using the KAPA Hyper Prep kit for Illumina platforms (Kapa Biosystems, Wilmington, MA, USA) according to manufacturer's directions with 9 cycles of library amplification. Illumina® TruSeq™ adapters were used enabling combination and simultaneous sequencing of 24 DNA samples using the MiSeq 300-cycle v2 reagent kit (Illumina). Raw data was aligned to a bespoke fasta file containing the reference sequences for all amplicons using the mem algorithm within BWA (http://bio-bwa.sourceforge.net/). The resultant SAM files were converted to binary form (BAM files) using SAMtools (http://samtools.sourceforge.net/) and variant calling was made with GATK's Unified Genotyper (v 2.8-1) [37] using default parameters with no downsampling.

2.3. Statistical analyses

2.3.1. Population analyses and haplotype inference

Genetic data obtained for 29 autosomal SNPs was tested for Hardy–Weinberg equilibrium (HWE) using Arlequin 3.5 software (http://cmpg.unibe.ch/software/arlequin35). HWE testing provided a necessary check for sequence variation, as deviations from HWE can occur from variants in PCR-primer sites leading to mistyping heterozygotes as homozygotes. However, HWE deviations can also be caused by population stratification, or other genetic factors in the study populations such as inbreeding, selection and genetic drift [38]. Lastly, the inference of SNP haplotypes of linked variants is more applicable than analysis of single SNPs when analyzing arrays of several SNPs in short chromosome segments [39]. Haploview version 4.2 (www.broadinstitute.org) was used to test the degree of linkage disequilibrium (LD) between four closely sited SNPs in *TCHH* (rs17646946, rs11803731, rs4845418 and rs12130862), and statistical haplotype reconstruction for these SNPs was performed using PHASE v. 2.1 software (http://stephenslab.uchicago.edu/software.html). All genomic positions used in LD analysis and haplotype inference were retrieved from the GRCh37.p13 genome build.

2.3.2. Association testing

The candidate SNPs were tested for association with hair morphology by using the binary logistic regression method with IBM SPSS Statistics v. 22 (SPSS Inc., Chicago, IL, USA). For this purpose hair morphology was categorized as straight vs. non-straight (straight vs. curly/wavy) due to low sample numbers for the curly hair category. Straight hair was coded as '1' vs. curly/wavy hair as '0' and association analyses used the Polish sample set of 333 individuals with straight hair and 195 with curly/wavy hair (75/195 and 120/195 respectively). Detailed characteristics of the Polish population sample set are given in Supplementary Table S2 and Fig.

1. Firstly, univariate association analysis for the six key hair morphology SNPs (rs17646946, rs11803731, rs4845418, rs12130862, rs1268789, rs7349332) was made followed by multivariate regression analysis involving simultaneous testing of the six SNPs as well as inferred *TCHH* haplotypes. Allelic odds ratios (ORs) with 95% confidence intervals (CIs) and respective P values were assessed for minor and/or major effect alleles (alleles associated with straight hair) categorized in an additive manner. Haplotypes in *TCHH* with frequencies exceeding 0.5% were included in association analyses considering an additive mode of inheritance. Lastly, the combined effect of applying just the three most strongly associated SNPs selected from multivariate regression analysis, was assessed by estimating the Genotype Risk Score (GRS). The GRS was calculated using a weighted risk allele counting approach, which sums the β parameter values of weighted risk alleles (from logistic regression) for the selected SNPs [35,40]. The resulting GRS was then tested for association with hair morphology estimating odds ratios, 95% CI and P values. In the next step, univariate association analyses were performed for the remaining 44 SNPs previously identified to be associated with MPB. The proportion of total variance in hair morphology explained by the tested variables was estimated using the Nagelkerke pseudo-R^2 statistic (IBM SPSS Statistics v. 22).

2.3.3. Prediction modeling

The predictive performance of the studied SNPs was evaluated during the development of prediction models using three approaches: logistic regression, neural networks and classification and regression trees (CRTs). Prediction models were constructed using the full set of 528 Polish samples and were built from the SNPs rs11803731 (*TCHH*); rs1268789 (*FRAS1*) and rs7349332 (*WNT10A*), selected as the most strongly associated from multivariate regression analysis. Models were tested using 10-fold cross validation, a procedure described in detail in [41]. Briefly, this method splits data into ten '*k*' portions of equivalent number. For each *k* (*k=1,2,...,10*) the *k*th part is excluded and the model built using the data of the other *k-1* parts. Then the prediction error is calculated on the excluded *k*th part of the data. The final prediction error is estimated by the mean of errors of 10 models built in the cross-validation procedure. The binary logistic regression model was estimated using block entry of variables (entry value 0.05). In the case of neural networks analysis, Multilayer Perceptron (MLP) was used with one hidden layer and an automatically selected number of neurons. The activation functions were hyperbolic

tangent for the hidden layer and softmax for the output layer. Lastly, the CRT algorithm was used with Gini impurity measure, the maximum tree depth of 5, the minimal number of cases in the parents nodes equal to 2 and in the child nodes equal to 1. To avoid over-fitting, trees were pruned with maximum difference in risk equal to 0.25 of standard error. Detailed descriptions of statistical methods used in this study are outlined in [41] for CRTs, in [42] for logistic regression and in [43] for neural networks. To assess the performance of the developed prediction models the summary statistics estimated area under the ROC curve (AUC), sensitivity, specificity and total number of correct calls according to several previous studies of forensic EVC-predictive tests [10, 13, 44-45]. The final prediction models, developed from the full Polish sample set, were then tested using an independent set of 142 samples from six European populations (Supplementary Table S3). All prediction analyses and resulting prediction models were developed using IBM SPSS statistics v.22 software.

3. Results

3.1. Population analyses and haplotype inference

Population analyses and haplotype inferences used 528 Polish individuals comprising 218 females and 310 males. The frequency of straight hair was 62.9% in males and 63.3% in females (Fig. 1). No significant difference in hair morphology was detected between males and females (P-value=0.998) and no significant impact of age was observed (P-value=0.403, Supplementary Table S2). Hardy-Weinberg equilibrium analysis of 29 autosomal SNPs indicated minor deviation for rs2942168, rs12373124 and rs1800547 located on a 337 kilobase chromosome 17 segment, giving P-value>0.035 although the effect of multiple hypothesis testing means this value can be discounted (Supplementary Table S4). SNPs rs12130862, rs17646946, rs11803731 and rs4845418 in or near *TCHH* showed strong linkage disequilibrium (r^2>0.75, with r^2>0.9 between the first three). The distribution of LD estimated by Haploview is shown in Supplementary Fig. 1. Reconstruction of haplotypes for these four SNPs in one LD block revealed two very common haplotypes: TGAG and AATC, with frequencies of 72.3% and 22.0%, respectively. An additional haplotype AATG (3.9% frequency) and five rare haplotypes (<1%) were reconstructed from the Polish sample (Table 1).

3.2. Association testing

Our replication of the GWAS results reported by Medland et al. [26] centered on univariate association analyses with binary logistic regression for straight vs. wavy/curly hair. Results confirmed the statistical significance of all six SNPs that Medland's study originally identified [26]. The highest significance (P-value=9.77×10^{-6}) was recorded for rs11803731 (*TCHH*) with an odds ratio (OR) of 2.0 (95% CI=1.5-2.7). According to the Nagelkerke R^2 statistic, this polymorphism explains 5.4% of the hair morphology variation in the studied population. Similar results were obtained for rs12130862, rs17646946 and rs4845418 forming a single LD block with rs11803731 (Table 2). Analysis of the inferred haplotypes showed association of the most frequent TGAG haplotype (72.3%) with curly/wavy hair and the AATC haplotype (22.0%) with straight hair. However, these haplotypes did not show stronger associations to hair morphology compared to rs11803731 genotypes alone. The strongest effect obtained for the AATC haplotype of OR=2.1 (1.5-2.9), P=1.87×10^{-5} and R^2 Nagelkerke=5.1%, was similar to the effect of rs11803731 by itself (Table 1). Association values recorded for rs7349332 (*WNT10A*) and rs1268789 (*FRAS1*) were found to be lower than those for SNPs in *TCHH* with OR=1.6 (95% CI=1.1-2.4, P-value=0.018) and OR=1.4 (95% CI=1.1-1.8, P-value=0.013) respectively. The fraction of hair morphology variation explained by these two SNPs reached just 1.4% and 1.6% respectively (Table 2).

Multivariate association analysis, including simultaneous analysis of all six SNPs and *TCHH* haplotypes confirmed association with hair morphology for rs11803731 in *TCHH*, rs7349332 in *WNT10A* and rs1268789 in *FRAS1*, with the strongest effect noted for rs11803731 (OR=2.0, 95% CI=1.5-2.8, P-value=1.12×10^{-5}) and weaker effects revealed for rs7349332 (OR=1.7, 95% CI=1.1-2.5, P-value=0.015) and rs1268789 (OR=1.4, 95% CI=1.0-1.8, P-value=0.022). This multivariate regression model gave χ^2 significance of P=3.85×10^{-7} and explains 8.2% of the total variation in hair morphology (Table 3). The combined effect of these 3 polymorphisms was further evaluated by estimating the genotype risk score (GRS). GRS was calculated as the weighted number of alleles associated with straight hair and was found to be associated with hair morphology with OR=2.7, 95% CI=1.9-3.9 and P-value=6.64×10^{-8}.

3.3. Investigation of SNPs previously associated with MPB

Analysis of 44 SNPs previously associated with MPB (androgenetic alopecia) revealed a suggestive association with hair morphology for just SNP rs4679955 (3q25.1). The minor allele rs4679955-A (0.43 allele frequency) was found to increase the odds of straight hair by a factor of 1.4 (95% CI=1.1-1.8, P-value=0.017). From the Nagelkerke R^2 statistic, this SNP explains 1.5% of variation in hair morphology in the Polish population (Supplementary Table 1).

3.4. Prediction modeling

Three types of prediction models were developed and then compared for performance. The 10-fold cross validation procedure was applied to 528 Polish samples using logistic regression, neural networks and CRT prediction methods. The prediction models were created using genotype data for rs11803731 (*TCHH*), rs7349332 (*WNT10A*) and rs1268789 (*FRAS1*), the three key predictors selected in multivariate logistic regression. All three prediction models showed high sensitivity of straight hair prediction and significantly lower levels of straight hair prediction specificity (equivalent to sensitivity of curly/wavy hair prediction) as shown in Table 6. The highest sensitivity of straight hair prediction was obtained with logistic regression (93.2%) and a similar value was obtained with neural networks (91.2%), while noticeably lower with CRT (77.3%). In contrast, the highest specificity of straight hair prediction was obtained with CRT at 34.8%, more than twice that of logistic regression (15.4%). Much lower straight hair prediction specificity compared to sensitivity signifies a reduced ability of the model to detect curly/wavy hair - classified as straight in a significant proportion of samples. Neural networks provided the best balance between sensitivity and specificity (straight hair sensitivity 91.2%, specificity 23.0%) and gave the highest number of total correct predictions (66.9%) of all the methods compared (Table 4). Moreover, overall accuracy of prediction measured by the area under the ROC curve parameter was highest for neural networks with an AUC=0.688 (Fig. 2, Table 4).

Assessing the performance of the developed prediction models using the probability threshold of 50% and the test set of 142 samples confirmed high sensitivity and low specificity of straight hair prediction (Table 5). The highest proportion of correct predictions was obtained with logistic regression (84.5%), while this success rate dropped with neural networks (64.1%) and CRT (61.3%). Due to the fact that the test set had a much higher ratio of straight (128) to curly/wavy hair samples (14) these values confirmed the results obtained for the Polish samples

indicating the highest ability of logistic regression to detect straight hair. In contrast, detailed analysis confirmed a much higher ability of neural networks and CRT methods to detect curly/wavy hair as measured by the specificity of straight hair prediction in each case: logistic regression=14.3%, neural networks=35.7% and CRT=50.0%. Predictive success rates in each European population are given in Supplementary Table S5. The proportion of correct predictions differed noticeably amongst these population samples and ranged from 54.6% in Germany to 91.3% in Netherlands using logistic regression method, 54.6% in Germany to 68.4% in Italy using neural networks and 56.5% in Netherlands to 77.8% in Spain using CRT. This effect may reflect certain inter-population differences in the genetics of hair morphology, however they can also arise from the very small sample sizes available for this study.

To better describe the predictive accuracy of each model, different probability thresholds were applied in the next stage of analysis. Moving from an initial 50% probability threshold, we assessed prediction success obtained with 60% and 65% thresholds, where probabilities below these values were defined as 'inconclusive results'. The number of inconclusive results was ~40% for the 60% probability threshold, but exceeded 70% in logistic regression and CRT applying the 65% threshold (Table 5). However, the proportion of correct predictions increased, particularly for neural networks and CRT reaching 78.7% and 94.9%, respectively, with 65% threshold. Notably, the neural network-65% threshold approach gave 50.0% specificity of straight hair prediction, while maintaining high level of sensitivity (81.4%) of straight hair prediction (Table 5).

A high proportion of inconclusive results (66.9% at 65% thresholds for neural networks) shows that the available SNP predictors are only informative in a small number of cases. Therefore, in the next step we performed detailed analysis of genotype combinations for three selected loci and genotype combinations giving the highest confidence of prediction were selected. Four genotype combinations with frequencies between 0.76% and 13.26% were found to provide >70% probability of straight hair (Fig. 3). The highest confidence of prediction was found with the TTGGGG combination (all six straight-hair associated alleles in rs11803731, rs7349332 and rs1268789; 4.5% frequency in Poland) giving 82-88% probabilities of straight hair, depending on the prediction model type. In 528 individuals from Poland, the TTGGGG genotype was present in 24: 20 phenotyped as straight-haired and two each as wavy- or curly-haired. In the 142 test-set, this genotype was observed just once and classified correctly as

straight-haired. In comparison, analyses of genotype combinations associated with the highest probabilities for curly/wavy hair were ambiguous. No combination of all 6 alleles associated with curly/wavy hair (AAAAAA) was recorded, while 5 associated alleles were at insufficient frequencies to permit proper interpretation (data not present).

4. Discussion

Hair morphology is a diverse and distinctive feature of human appearance and therefore prediction of individual hair type based on evidential DNA has obvious informative value for investigative purposes. Straight hair is considered to be the derived variant of hair morphology that most likely evolved in parallel in Europe and Asia [23-25]. The prevalence of straight hair in European populations has been estimated to be ~45-50% [14-15] below the proportion of ~63% in the Polish study sample. This difference may relate to some inter-population variation seen in Europeans. It has also been suggested that males are ~5% more likely to have straight hair than females, and additionally, curliness of hair in males can increase slightly with age, possibly due to hormonal regulation. However, this effect was not observed in females [26]. In our study, no effect of age and gender on hair morphology was found.

Although hair morphology is evidently highly heritable, very little is currently known about its genetic determination, while just one published GWA study indicated the single gene of *TCHH* reached genome-wide significance [26]. Our study provides additional evidence for association, confirming that the four tested polymorphisms in or near *TCHH* are in strong LD (r^2>0.75). Eriksson's study demonstrated that each rs17646946-A allele reduces the degree of hair curliness by about 0.29 points on a scale from 0 to 5 [27]. In our study the strongest effect was shown for rs11803731-T increasing the odds of straight hair by a factor of OR≈2.0. According to the results of multivariate association analysis, three other *TCHH* SNPs did not offer additional information on hair morphology determination, in line with the results of Medland et al. [26]. Among the four SNPs analyzed in the 1q21 region, rs11803731 appears to be the most likely functional variant, although this suggestion needs to be confirmed in additional studies. SNP rs11803731 is a non-synonymous variant in the third exon of *TCHH* and produces a methionine to leucine substitution at codon 790. Interestingly, in silico functional annotation using various programs indicates a regulatory rather than a structural effect for this polymorphism [26]. Medland et al. also reported suggestive association with hair morphology for *WNT10A* and *FRAS1* [26]. Association of the

intronic rs7349332 SNP in *WNT10A* with natural variation in hair morphology was also shown by Eriksson et al. [27]. Our study indicates the associations between hair morphology and SNPs rs7349332 in *WNT10A* and rs1268789 in *FRAS1* are weaker than that found for rs11803731 in *TCHH*. The genotype risk score evaluating these three SNP's combined effect, gave a straight hair odds ratio of 2.7, while we estimate that the three loci explain ~8.2% of the total variance in hair morphology in the Polish sample. Analysis of variation in *TCHH* alone was estimated by Medland et al. to explain ~6% of hair morphology variation in Europeans [26].

Our evaluations of the predictive performance of the three SNPs with three methods and 10-fold cross validation revealed higher capacity to predict straight hair compared to wavy/curly hair, producing high sensitivity and low specificity of straight hair prediction. The highest predictive values were obtained with neural networks with an overall accuracy of AUC=0.688 and total correct predictions of 66.9%. This method was found to be more accurate than logistic regression which is a commonly chosen method for forensic EVC prediction [11, 13, 44, 46-47]. Neural networks belong to the pattern recognition methods that focus on identification of patterns and regularities in complex data and can be considered as worthwhile alternatives to the traditional methods such as logistic regression [43, 48]. Neural networks imitate the neuronal connections in the human brain in a highly simplified model [49] consisting of nodes arranged in one input, several hidden and a single output layers that are all connected. Particular connections are assigned appropriate weights which finally describes the architecture of the optimum neural networks. Neural networks also gave the best results when the prediction models were evaluated using the test set of samples from six European populations. Straight hair sensitivity predictive success ranged from 67.2-89.5%, depending on probability thresholds applied, but these values were lower than those of the 10-fold cross validation used to test the Polish sample set. This difference may be related to different phenotyping regimes used in the case of the Polish samples (assessed by a dermatology specialist combined with interview) compared to the samples from other European populations (in most cases assessed from photographs). Classification of hair morphology from photographs alone can result in a certain level of error, especially in the case of short-haired individuals where distinguishing between straight and wavy hair may be problematic. The application of neural networks was found to be more sensitive in detecting wavy/curly hair (25-50% depending on the probability threshold) in the European test set compared to logistic regression method (0-25%), confirming the results obtained for the Polish

samples. The best results were obtained with neural networks applying a 65% probability threshold. The application of probability thresholds is often used in forensic EVC prediction studies [6-7, 13, 50], as it improves the confidence of prediction. In our study, despite the high level of non-classification at the 65% threshold, the correct call rate increased substantially from 64.1% to 78.7%.

Based on our study of three SNPs explaining only ~8% of variation in hair morphology, forensic prediction of this trait should just be made from particular genotype combinations that provide very high probabilities and confidence of prediction. The data indicates that the best straight hair predictor is the TTGGGG genotype (rs11803731, rs7349332 and rs1268789) containing all six straight hair-associated alleles. This combined genotype gives >80% probability of straight hair. Additional research is the critical next step needed to detect additional more weakly associated markers involved in human hair morphology. Analysis of gene-gene interactions will also be informative, as the role of epistatic effects is often emphasized in the determination of complex traits and such analyses benefitted studies of human pigmentation traits [7,47,51]. Recently, a possible link between genetic determination of hair loss and hair morphology has been suggested by Heilmann et al. [28]. This study demonstrated the association of rs7349332 in *WNT10A* with MPB. Our previous study of MPB in 305 males from 5 European populations failed to detect an association, although a positive impact on the prediction of MPB was noted [13]. In the present study, a suggestive association of rs4679955 on 3q25.1 with hair morphology was detected. This SNP was previously found to be associated with androgenetic alopecia [28] and the effect observed should be investigated further on a larger sample. SNP rs4679955 is located between *SUCNR1*; a gene encoding succinate receptor 1, and *MBNL1*; a gene encoding muscleblind-like splicing regulator 1. The role of rs4679955 in the determination of hair morphology and male hair distribution as well as assessment of this SNP's prediction capacity for both of these FDP traits will merit further studies.

5. Conclusion

Studies searching for the genetic determinants of human hair morphology can begin to offer a complementary physical trait to the set of EVCs already established for FDP. This first pilot study based on three main hair morphology-associated SNPs in *TCHH*, *WNT10A* and *FRAS1* found only 8.2% of the trait's variance is explained by these loci. Their genotypes provide a high level of straight hair prediction sensitivity but a much lower level of specificity. Assessments of three prediction methods indicated neural networks provide the best balance between sensitivity and specificity compared to logistic regression. Therefore, we suggest this prediction method is important to consider in the future development of forensic EVC prediction regimes. From the results of our study the TTGGGG genotype of rs11803731, rs7349332, rs1268789 is a good predictor of straight hair and gives >80% probability of straight hair. However, since this genotype is observed in only 4.5% of individuals, more studies of hair morphology are needed in order to select additional markers underlying this trait and to improve its prediction accuracy for forensic application.

Acknowledgements

The work leading to these results was financially supported from a grant from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE). The authors thank all sample donors for their contribution to this project.

References

1. M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic. Sci. Int. Genet. 3 (2009) 154–161.
2. W. Branicki, Studies on predicting pigmentation phenotype for forensic purposes, Probl. Forensic Sci. 77 (2009) 29–52.
3. M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.
4. W. Branicki, E. Pośpiech, T. Kupiec, J. Styrna, A new dimension of the forensic DNA expertise - the need for training experts and expertise recipients, Arch. Med. Sadowej Kryminol. 64 (2014) 175-194.
5. M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. (2015) doi: 10.1016/j.fsigen.2015.02.003.
6. S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, et al., IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic. Sci. Int. Genet. 5 (2011) 170–180.
7. Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. de Cal Casares, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2013) 28–40.
8. J.S. Allwood, S. Harbison, SNP model development for the prediction of eye colour in New Zealand, Forensic Sci. Int. Genet. 7 (2013) 444–452.
9. K.L. Hart, S.L. Kimura, V. Mushailov, Z.M. Budimlija, M. Prinz, et al., Improved eye- and skin-color prediction based on 8 SNPs, Croat. Med. J. 54 (2013) 248–256.
10. W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech et al., Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443–454.
11. S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, et al., The HIrisplex System for simultaneous prediction of hair and eye colour categories including hair colour shade from DNA, Forensic Sci. Int. Genet. 7 (2013) 98–115.

12. J. Söchtig, C. Phillips, O. Maroñas, A. Gómez-Tato, R. Cruz, J. Alvarez-Dios et al., Exploration of SNP variants affecting hair colour prediction in Europeans, Int J Legal Med (2015), doi:10.1007/s00414-015-1226-y.
13. M. Marcińska, E. Pośpiech, S. Abidi, J. Dyrberg Andersen, M. van den Berge, Á. Carracedo, M. Eduardoff, A. Marczakiewicz-Lustig, N. Morling, T. Sijen, M. Skowron, J. Söchtig, D. Syndercombe-Court, N. Weiler, The EUROFORGEN-NoE Consortium; D. Ballard, C. Børsting, W. Parson, C. Phillips, W. Branicki, Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness, PLoS One (2015) doi: 10.1371/journal.pone.0127852.
14. A. Franbourg, P. Hallegot, F. Baltenneck, C. Toutain, F. Leroy, Current research on ethnic hair, J. Am. Acad. Dermatol. 48 (2003) S115-119.
15. G. Loussouarn, A.L. Garcel, I. Lozano, C. Collaudin, C. Porter, S. Panhard, D. Saint-Léger, R. de La Mettrie, Worldwide diversity of hair curliness: a new method of assessment, Int. J. Dermatol. 46 (2007) 2-6.
16. S.E. Medland, G. Zhu, N.G. Martin, Estimating the heritability of hair curliness in twins of European ancestry. Twin. Res. Hum. Genet. 12 (2009) 514-8.
17. E.J. Van Scott, T.M. Ekel, Geometric relationships between the matrix of the hair bulb and its dermal papilla in normal and alopecic scalp, J. Invest. Dermatol. 5 (1958) 281-287.
18. S. Thibaut, B.A. Bernard, The biology of hair shape, Int. J. Dermatol. 44 (2005) 2-3.
19. S. Thibaut, P. Barbarat, F. Leroy, B.A. Bernard, Human hair keratin network and curvature, Int. J. Dermatol. 46 (2007) 7-10.
20. P.M. Steinert, D.A. Parry, L.N. Marekov, Trichohyalin mechanically strengthens the hair follicle: multiple cross-bridging roles in the inner root sheath, J. Biol. Chem. 278 (2003) 41409-41419.
21. S. Yamamoto, K. Hirai, Y. Hasegawa-Oka, Y. Hirai, Molecular elements of the regulatory control of keratin filament modulator AHF/trichohyalin in the hair follicle, Exp. Dermatol. 18 (2009) 152-159.
22. G.E. Westgate, N.V. Botchkareva, D.J. Tobin, The biology of hair diversity, Int. J. Cosmet. Sci. 35 (2013) 329-336.

23. A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M. S. Mustofa, U. Samakkarn, W. Settheetham-Ishida, T. Ishida, Y. Morishita, T. Furusawa, M. Nakazawa, R. Ohtsuka, K. Tokunaga, A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness, Hum. Mol. Genet. 17 (2008) 835-843.
24. A. Fujimoto, N. Nishida, R. Kimura, T. Miyagawa, R. Yuliwulandari, L. Batubara, M. S. Mustofa, U. Samakkarn, W. Settheetham-Ishida, T. Ishida, Y. Morishita, T. Tsunoda, K. Tokunaga, J. Ohashi, FGFR2 is associated with hair thickness in Asian populations, J. Hum. Genet. (54) (2009) 461-5.
25. J. Tan, Y. Yang, K. Tang, P.C. Sabeti, L. Jin, S. Wang, The adaptive variant EDARV370A is associated with straight hair in East Asians, Hum. Genet. 132 (2013) 1187-1191.
26. S.E. Medland, D.R. Nyholt, J.N. Painter, B.P. McEvoy, A.F. McRae, G. Zhu, S.D. Gordon, M.A. Ferreira, M.J. Wright, A.K. Henders, M.J. Campbell, D.L. Duffy, N.K. Hansell, S. Macgregor, W.S. Slutske, A.C. Heath, G.W. Montgomery, N.G. Martin, Common variants in the trichohyalin gene are associated with straight hair in Europeans, Am. J. Hum. Genet. 85 (2009) 750-755.
27. N. Eriksson, J.M. Macpherson, J. Y. Tung, L.S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, I. Pe'er, J. Mountain, Web-based, participant-driven studies yield novel genetic associations for common traits, PLoS Genet. 6 (2010) e1000993.
28. S. Heilmann, A.K. Kiefer, N. Fricker, D. Drichel, A.M. Hillmer, C. Herold et al., Androgenetic alopecia: Identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology, J. Invest. Dermatol. 133 (2013) 1489-1496.
29. D.A. Prodi, N. Pirastu, G. Maninchedda, A. Sassu, A. Picciau, M.A. Palmas, et al., EDA2R is associated with androgenetic alopecia, J. Invest. Dermatol. 128 (2008) 2268–2270.
30. J.B. Richards, X. Yuan, F. Geller, D. Waterworth, V. Bataille, D. Glass, et al., Male pattern baldness susceptibility locus at 20p11, Nat. Genet. 40 (2008) 1282–1284.
31. A.M. Hillmer, F.F. Brockschmidt, S. Hanneken, S. Eigelshoven, M. Steffens, A. Flaquer, et al., Susceptibility variants for male-pattern baldness on chromosome 20p11, Nat. Genet. 40 (2008) 1279–1281.

32. A. Hillmer, J. Freudenberg, S. Myles, S. Herms, K. Tang, D.A. Hughes, et al. Recent positive selection of a human androgen receptor/ectodysplasin A2 receptor haplotype and its relationship to male pattern baldness, Hum. Genet. 126 (2009) 255–264.
33. J.E. Cobb, S.G. Zaloumis, K.J. Scurrah, S. B. Harrap, J.A. Ellis. Evidence for two independent functional variant for androgenetic alopecia around the androgen receptor gene, Exp. Dermatol. 19 (2010) 1026-1028.
34. F.F. Brockschmidt, S. Heilmann, J.A. Ellis, S. Eigelshoven, S. Hanneken, C. Herold, et al., Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness, Br. J. Dermatol. 165 (2011) 1293-1302.
35. R. Li, F.F. Brockschmidt, A.K. Kiefer, H. Stefansson, D.R. Nyholt, K. Song, et al. Six novel susceptibility loci for early-onset androgenetic alopecia and their unexpected association with common diseases, PLoS Genet. 8 (2012) e1002746.
36. S. Redler, F.F. Brockschmidt, R. Tazi-Ahnini, D. Drichel, M.P. Birch, K. Dobson, et al. Investigation of the male pattern baldness major genetic susceptibility loci *AR*/*EDA2R* and 20p11 in female pattern hair loss, Br. J. Dermatol. 166 (2012) 1314-1318.
37. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297-303.
38. D.J. Balding, A tutorial on statistical methods for population association studies, Nat. Rev. Genet. 7 (2006) 781-791.
39. W. Barendse, Haplotype Analysis Improved Evidence for Candidate Genes for Intramuscular Fat Percentage from a Genome Wide Association Study of Cattle, PLoS One (2011) doi:10.1371/journal.pone.0029601.
40. M. Kirin, A. Chandra, D.G. Charteris, C. Hayward, S. Campbell, I. Celap, et al. Genome-wide association study identifies genetic risk underlying primary rhegmatogenous retinal detachment, Hum. Mol. Genet. (22) (2013) 3174-3185.
41. T. Hastie, T. Trevor, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition, Springer Series of Statistics, (2009).
42. D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied logistic regression. Third Edition, Hoboken N: John Wiley & Sons (2013).

43. R. Rojas, Neural Networks: a systematic introduction. Springer. Berlin (1996).
44. F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, et al., Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (2009) R192–R193.
45. E. Pośpiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Prediction of eye color from genetic data using Bayesian approach, J. Forensic Sci. 57 (2012) 880-886.
46. V. Kastelic and K. Drobnic. A single-nucleotide polymorphism (SNP) multiplex system: the association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model, Croat. Med. J. 53 (2012) 401-8.
47. E. Pośpiech, A. Wojas-Pelc, S. Walsh, F. Liu, H. Maeda, T. Ishikawa, M. Skowron, M. Kayser, W. Branicki, The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction, Forensic Sci. Int. Genet. 11 (2014) 64-72.
48. M. Schumacher, Rossner, R. Neural networks and logistic regression: Part I Comp. Stats. & Data Anal. 21 (1996) 661–682.
49. L. Tarassenko, A guide to neural computing applications, London: Arnold Publishers, (1998), pp. 1-4.
50. V. Kastelic, E. Pośpiech, J. Draus-Barini, W. Branicki, K. Drobnič. Prediction of eye color in the Slovenian population using the IrisPlex SNPs. Croat. Med. J. 54 (2013) 381-386.
51. E. Pośpiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Gene-gene interactions contribute to eye colour variation in humans, J. Hum. Genet. 56 (2011) 447-455.

Figure 1. Hair morphology distribution of curly, wavy and straight hair in the Polish study sample used for the association studies.

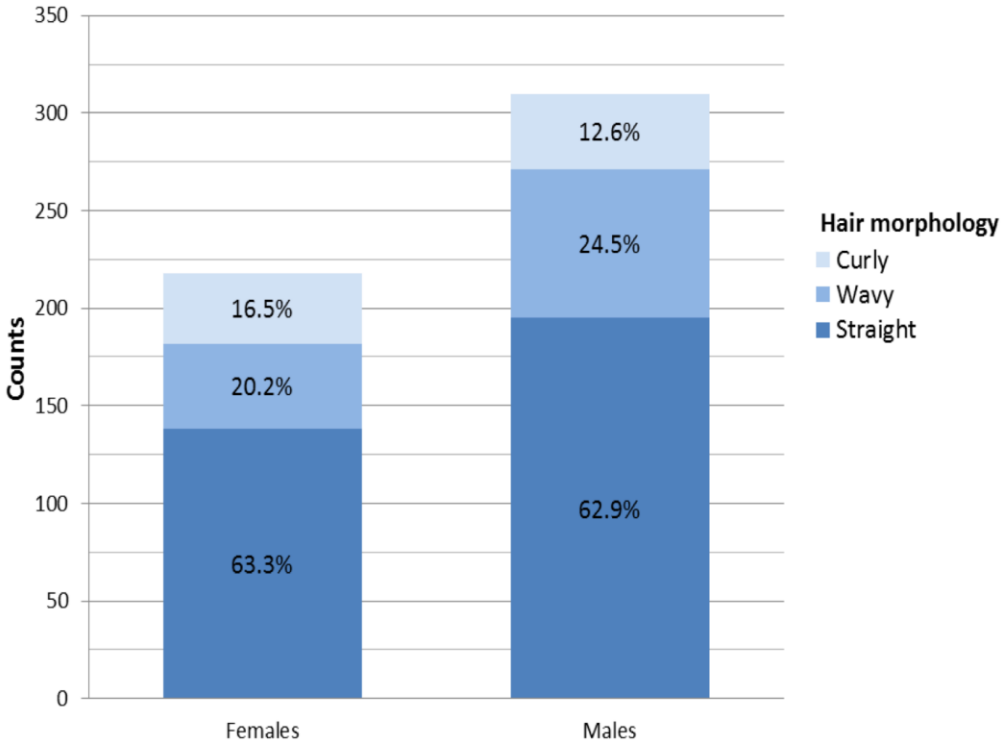

Figure 2. ROC curve chart for straight vs. curly/wavy hair prediction using three prediction models based on 528 Polish samples. Area under the ROC curve illustrates the overall accuracy of prediction measured by sensitivity and specificity assigned at various probability thresholds. AUC=0.5 means random prediction and AUC=1.0 means perfect prediction.

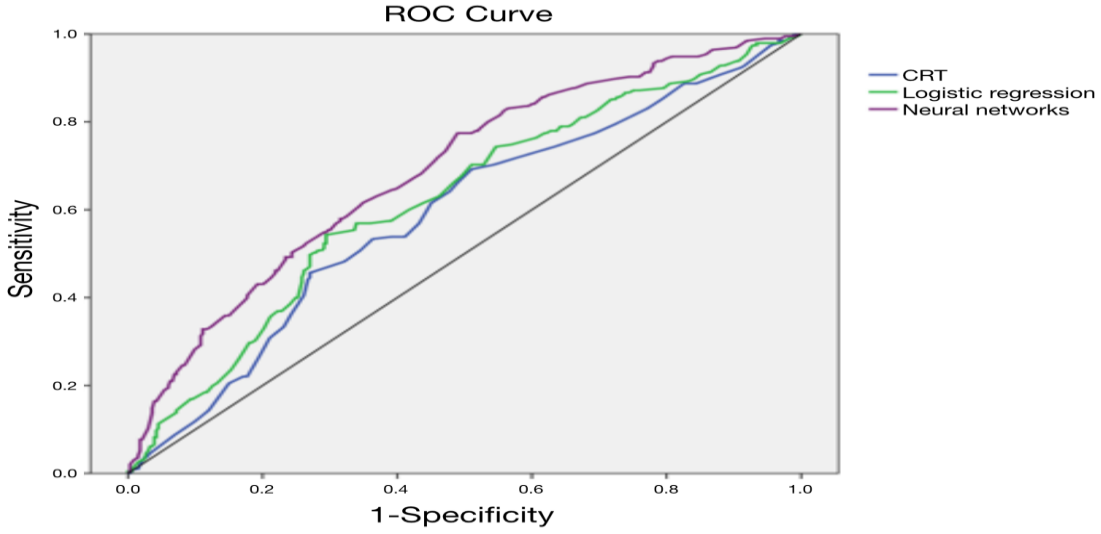

Figure 3. Genotype combinations of SNPs rs11803731, rs7349332, rs1268789 that give ≥70% probability for straight hair and their frequencies.

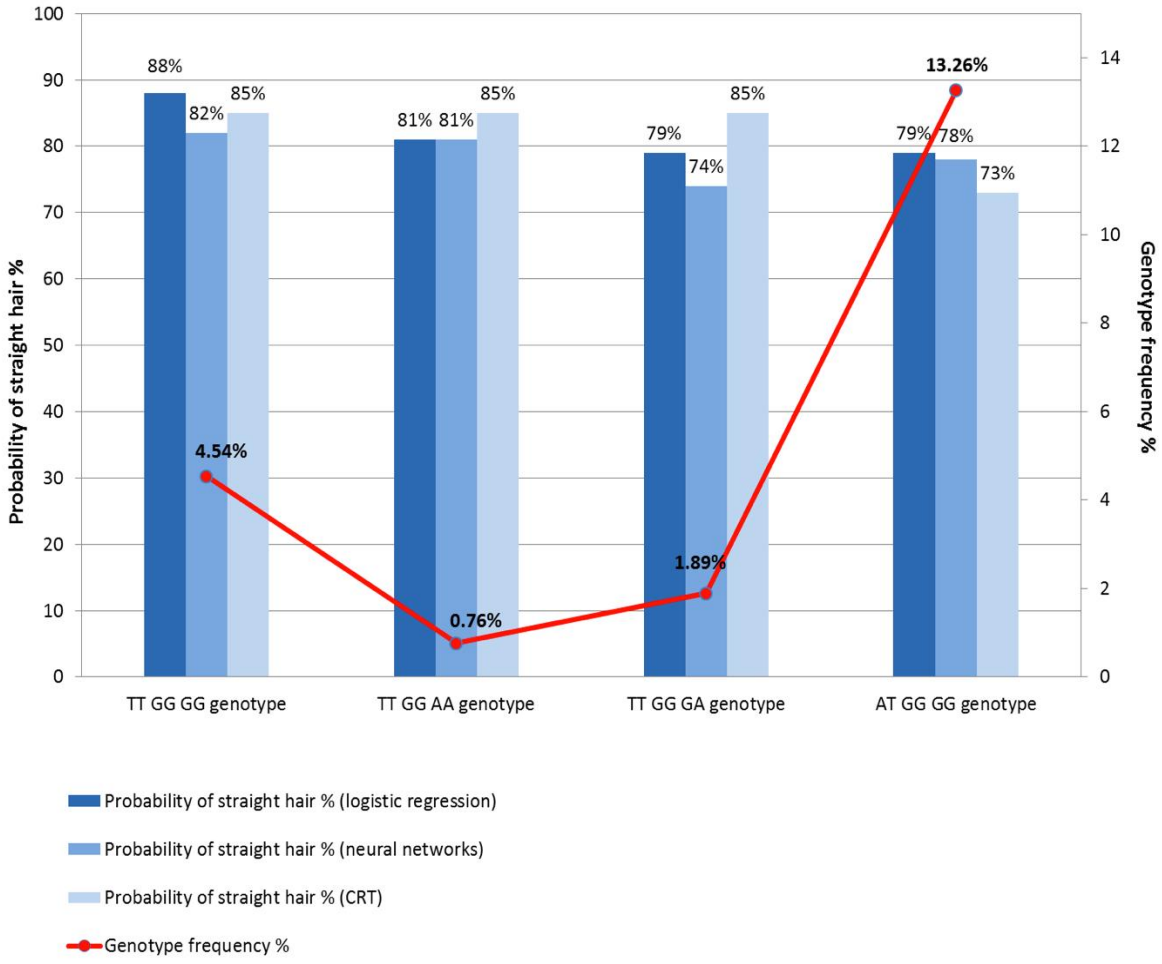

Table 1. Inferred haplotypes and their association with hair morphology obtained from the Polish population sample. Bold values indicate the two most strongly associated haplotypes to straight hair.

Haplotype analysis for rs12130862, rs17646946, rs11803731, rs4845418 (*TCHH*)					
No.	Haplotype	Frequency (%)	Association testing for straight vs. wavy/curly hair		
			OR (95% CI)	P-value	Nagelkerke R^2 statistic
1	**TGAG**	**764 (72.3)**	**0.6 (0.4-0.8)**	**$3.21x10^{-4}$**	**3.5%**
2	TGAC	1 (0.1)	NT	NT	NT
3	TAAG	9 (0.9)	0.3 (0.1-1.2)	0.079	NT
4	TATC	2 (0.2)	NT	NT	NT
5	AGAG	4 (0.4)	NT	NT	NT
6	AAAG	2 (0.2)	NT	NT	NT
7	AATG	41 (3.9)	1.3 (0.6-2.5)	0.471	NT
8	**AATC**	**233 (22.0)**	**2.1 (1.5-2.9)**	**$1.87x10^{-5}$**	**5.1%**
Total:		1056 (100.0)	-	-	-

OR: Odds ratios estimated for straight hair
NT: Not tested

Table 2. Univariate association analysis results for six hair morphology predictive SNPs.

SNP	Gene	Chr	Chr position GRCh37.p13	Alleles	MAF	Effect allele*	Univariate association analysis		
							OR (95% CI) for the effect allele	P-value	Nagelkerke R^2
rs12130862	*Near TCHH*	1	152027015	A/T	A 0.27	A	1.8 (1.4-2.5)	$7.73x10^{-5}$	4.3%
rs17646946	*TCHHL1*	1	152062767	G/A	A 0.27	A	1.8 (1.3-2.4)	$1.02x10^{-4}$	4.1%
rs11803731	*TCHH*	1	152083325	A/T	T 0.26	T	2.0 (1.5-2.7)	$9.77x10^{-6}$	5.4%
rs4845418	*Near TCHH*	1	152136230	C/G	C 0.23	C	2.1 (1.5-2.9)	$1.76x10^{-5}$	5.1%
rs7349332	*WNT10A*	2	219756383	G/A	A 0.11	G	1.6 (1.1-2.4)	0.018	1.4%
rs1268789	*FRAS1*	4	79280693	G/A	A 0.31	G	1.4 (1.1-1.8)	0.013	1.6%

* The allele associated with straight hair

Table 3. Multivariate association analysis results for hair morphology predictive SNPs.

SNP	Gene	Chr	Allele variants	Effect allele*	Multivariate association analysis	
					OR (95% CI) for the effect allele	P-value
rs11803731	*TCHH*	1	A/T	T	2.0 (1.5-2.8)	1.12×10^{-5}
rs7349332	*WNT10A*	2	G/A	G	1.7 (1.1-2.5)	0.015
rs1268789	*FRAS1*	4	G/A	G	1.4 (1.0-1.8)	0.022
P value of a model						3.85×10^{-7}
Nagelkerke R^2 statistic						8.2%
GRS** association						OR=2.7, 95% CI=1.9-3.9, $P=6.64 \times 10^{-8}$

* The allele associated with straight hair

** Combined Genotype Risk Score calculated for all three SNPs

Table 4. Parameters describing the accuracy of straight vs. curly/wavy hair prediction using SNPs: rs11803731; rs7349332; rs1268789; with three statistical approaches. Prediction parameters were assessed for 528 Polish samples using 10-fold cross validation. Light to dark cells indicate increasing sensitivity and specificity of straight hair prediction.

Hair morphology prediction of Polish sample set			
Parameters of prediction	Prediction model type		
	Logistic Regression	Neural Networks	CRT
AUC	0.622	0.688	0.589
Straight hair prediction sensitivity* %	93.2	91.2	77.3
Straight hair prediction specificity* %	15.4	23.0	34.8
Total number of correct calls %	63.8	66.9	60.8

*Sensitivity and specificity of straight hair prediction equate to specificity and sensitivity of curly/wavy hair prediction respectively.

AUC: Area Under the ROC Curve

CRT: Classification and Regression Tree

Table 5. Performance of the developed prediction models with the test set of 142 samples, assessing predictive success at three probability thresholds. Bold values indicate the model with the best predictive performance (i.e. the best balance between sensitivity and specificity)

Hair morphology prediction in an external testing set of 6 European populations					
Prediction model type	Prediction threshold	Straight hair prediction sensitivity* %	Straight hair prediction specificity* %	Inconclusive results %	Total number of correct calls %
Logistic regression	50%	92.2 (118/128)	14.3 (2/14)	-	84.5 (120/142)
	60%	87.3 (69/79)	25.0 (2/8)	38.7 (55/142)	81.6 (71/87)
	65%	90.0 (36/40)	0.0 (0/2)	70.4 (100/142)	85.7 (36/42)
Neural Networks	50%	67.2 (86/128)	35.7 (5/14)	-	64.1 (91/142)
	60%	89.5 (68/76)	25.0 (2/8)	40.9 (58/142)	83.3 (70/84)
	65%	**81.4 (35/43)**	**50.0 (2/4)**	**66.9 (95/142)**	**78.7 (37/47)**
CRT	50%	62.5 (80/128)	50.0 (7/14)	-	61.3 (87/142)
	60%	100.0 (80/80)	0.0 (0/7)	38.7 (55/142)	92.0 (80/87)
	65%	100.0 (37/37)	0.0 (0/2)	72.5 (103/142)	94.9 (37/39)

*Sensitivity and specificity of straight hair prediction correspond respectively to specificity and sensitivity of curly/wavy hair prediction.

CRT: Classification and Regression Tree